ADAPTIVE TESTING RESEARCH AT MINNESOTA— OVERVIEW, RECENT RESULTS AND FUTURE DIRECTIONS [1]

DAVID J. WEISS
University of Minnesota

Adaptive Testing and Error Reduction

The general objective of our research program on adaptive testing is to view it from a perspective which identifies several sources of potential error in test scores, and to study adaptive testing as a means for reducing these errors of measurement.

The first general source of error that we have been concerned with for some time is the error that results from the mismatch of item difficulties in an ability test with the individual's ability. Obviously, the testee's ability is not known at the start of testing. But the different strategies of adaptive testing that have been proposed can be viewed as different ways of matching item difficulties with testee ability and sequentially estimating the testee's ability. Consequently, one of our major focuses is to determine the best, or at least better, ways of adapting item difficulties to individual abilities.

We are approaching this in two complementary ways. First, we have been doing live computerized testing. Since late 1972 we have tested more than 5,000 subjects on a variety of strategies of adaptive testing. But live testing cannot provide the answer to all the questions concerning which strategies are best under which conditions, because there are too many questions to be answered. Therefore, we are using computer simulation to supplement and extend the results that we obtain from live testing.

Our general strategy is to implement an adaptive testing strategy in live testing to obtain some data with an arbitarily structured live adaptive test–data such as characteristics of score distributions and test-retest reliabilities. Then, our ultimate goal is to build a computer simulation model which will accurately reflect the results that we obtain from live testing. With the computer simulation model we can then very rapidly study different variations of the adaptive testing strategy. The next step is to verify the simulation results in live testing.

Thus far we have not yet developed a simulation model which completely reflects how live testees respond, but we are making progress toward that goal. The computer simulations are necessary because of the rapidity with which we can study various alternatives. The live testing is necessary, obviously, because it's people who take tests and not computers using hypothetical items or hypothetical subjects. So it is necessary to re-verify the results of the computer simulations to make sure that they still reflect what real people do given the variations we have made in the strategies studied in the simulations.

The second main focus of our research is a concern with the psychological effects of adaptive testing. Here we are concerned with identifying the psychological aspects of testing and the test environment which can introduce error into test scores. These variables include guessing, test anxiety, boredom, frustration, and racial or ethnic group effects.

Guessing can obviously artifically increase test scores; frustration, anxiety, motivation and other factors can result in test scores lower than true ability. All of these, therefore, are sources of error in test scores which are due to the psychological effects of testing.

We are also concerned with the psychological effects that will result from the man-machine interface. This, from our experience, is going to be an important problem in computerized adaptive testing. There are different kinds of computer systems on which we can implement adaptive testing and each of those computer systems has its positive and negative effects on testee behavior. There are different kinds of terminal devices for adaptive testing and each kind of terminal device displays in different ways and at different speeds. All of these variations in the man-machine interface are going to be new problems for us to consider in the years to come. Past research has demonstrated that answer sheets in paper and pencil testing sometimes had an effect on test scores. Similarly, research in adaptive testing will need to study different kinds of CRTs, different kinds of computer systems and different display speeds as part of the psychological effects of computerized testing.

A third source of error that we are concerned with has been briefly discussed this morning by Dr. Samejima; this is error that results from not extracting enough information from a testee's response to a test item. To date, most psychometric research has been concerned with binary or 0-1 scoring. But, as Dr. Samejima has indicated, we can get more information out of a test response if we treat it as a graded item. Our research extends that reasoning to continuous responses using the continuous case of latent trait theory. The continuous case is operationalized by probabilistic responding.

[1] Early development work on this research was supported during 1969 and 1970 by grants from the General Research Fund of the Graduate School, University of Minnesota. Research reported in this paper was supported since early 1972 by Personnel and Training Research Programs, Office of Naval Research, Contract No. N00014-67-A-0113-0029, NR 150-343. Special thanks are due to John DeWitt, our project programmer, without whom this research would have been almost impossible.

This aspect of our research is concerned with integrating probabilistic responding with adaptive testing. Probabilistic responding, like adaptive testing, can result in horizontal information functions. This implies that if we put adaptive testing and probabilistic responding together we will have extremely powerful methods of reducing errors in test scores due to the incomplete use of test responses.

The fourth source of error that we are studying is the error that results from deviations from unidimensionality. Latent trait theory, as it is usually used in testing, is based on the assumption of unidimensionality, although there are multidimensional latent trait models being developed. But dimensionality that is defined on a group, such as the unidimensionality of latent trait theory, does not necessarily hold true for an individual. That is dimensionality defined by factor analysis or other methods, when applied to an individual, assumes that the individual is the typical or average member of the group on which the dimensionality was defined. Thus, in the testing situation, when a set of "unidimensional" items is administered to an individual, the result may be a set of responses that are not unidimensionally determined.

Consequently, our research is concerned with individual-item pool interactions–the interaction of one individual with a set of "unidimensional" items. We are studying item response protocols of this nature to determine if meaningful deviations from unidimensionality do occur for specific individuals. If they do, we will then attempt to develop interactive testing models that will take account of intra-individual multidimensionality in an adaptive testing situation.

The focus of our research effort, as you can see, is with the *individual.* We are concerned with identifying those sources of error in test scores which result in the over- or under-estimation of *each individual's* ability.

Recent Results

Most of our recent results are concerned with the psychometric effects of adaptive testing, or the comparison of branching strategies. Thus far we have reported initial results from both live testing and computer simulation on a simple two-stage test (Betz & Weiss, 1973, 1974; Larkin & Weiss, 1975) and a pyramidal branching strategy (Larkin & Weiss, 1974, 1975). Below, I will report some results from a flexilevel test (Betz & Weiss, 1975) and some data on my stratified adaptive test (Weiss, 1975). Mr. McBride will present some data using Owen's (1975) Bayesian adaptive testing strategy.

In general, the findings that we have to date show that adaptive tests have higher test-retest stabilities–a very practical and useful criterion–when controlled for number of items and memory effects. Adaptive tests also tend to show, in simulation studies, better distributions of ability estimates. That is, ability estimates better reflect the distribution of generated ability. And, in general, adaptive tests give information functions which are less variable throughout the ability range, in support of Lord's theoretical findings (see Weiss & Betz, 1973).

Flexilevel ability testing. Figure 1 shows the item structure for Lord's (1971a,b) flexilevel test. In this testing strategy there is one item at each of a number of difficulty levels; item 19 is the most difficult item and item 18 the least difficult item. Everyone starts the flexilevel test with an item of median difficulty. Items with odd numbers increase in difficulty as they deviate from the median, and items with even numbers decrease in difficulty.

Figure 2 shows the paths taken by three different people through a ten-stage flexilevel test. Starting with the first item, a correct response leads to the next more difficult item which has not yet been administered. An incorrect response leads to the most difficult of the unadministered easier items. Figure 2a shows a high ability testee going through a flexilevel test, Figure 2b is for an average ability testee, and Figure 2c is for a low ability testee.

Our live-testing study of flexilevel testing (Betz & Weiss, 1975) used a flexilevel test in which each testee would answer 40 items, requiring a 79-item structure. That test and a conventional peaked paper-and-pencil type test, administered on a computer to control for novelty effects, was administered to 130 individuals. The same tests were then used in a computer simulation study. That study used 10,000 "subjects" sampled from a normal distribution of ability, and an additional 1600 subjects, 100 at each of 16 levels of ability. From these simulation data we calculated information functions, and test-retest or parallel forms reliability. From the live-testing study we calculated test-retest reliabilities, and other data describing score distributions.

The major result from the live-testing study was that flexilevel test scores were no more stable on retest than scores on the conventional test; test-retest stabilities for the two were virtually identical. The major result from the simulation study is shown in Figure 3, which displays information functions for the conventional and flexilevel tests. Figure 3 shows two findings which were not predicted by test theory.

First, test theory (e.g., Lord, 1971c) predicts that the conventional test will always result in higher levels of information, i.e., better measurement, than any adaptive test at the median of the ability distribution. Figure 3 shows that the flexilevel test had higher levels of the information function at the median (θ=0) of the ability distribution. The second prediction from test theory (Lord, 1971b) was that the flexilevel test should yield a relatively horizontal information function. Figure 3 shows an information function for the flexilevel test which is quite divergent from horizontal. In fact, the standard deviations of the information functions show that the flexilevel test had a larger standard deviation than did the conventional test; that means that the flexilevel test tended to be less equi-precise than the conventional test, at different levels of the ability distribution.

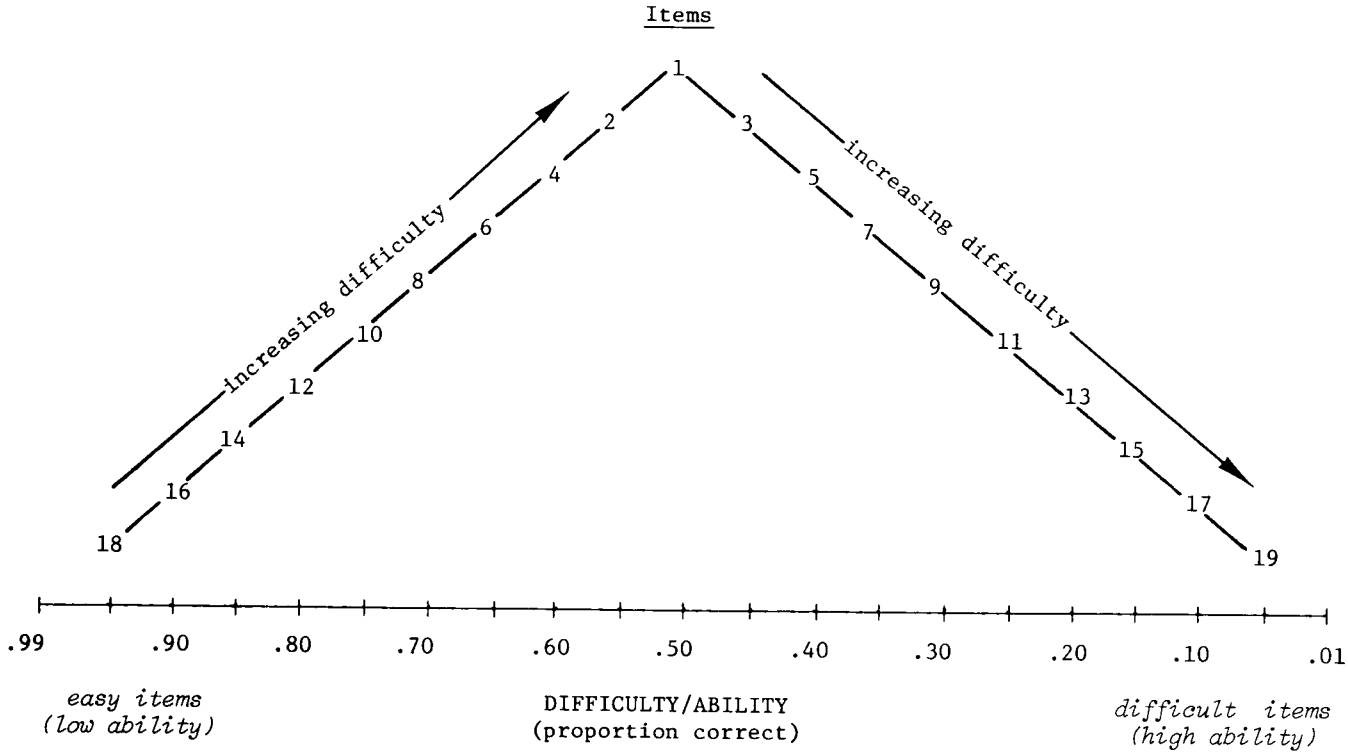


Figure 1

Items Structure for a Ten-stage Flexilevel Test

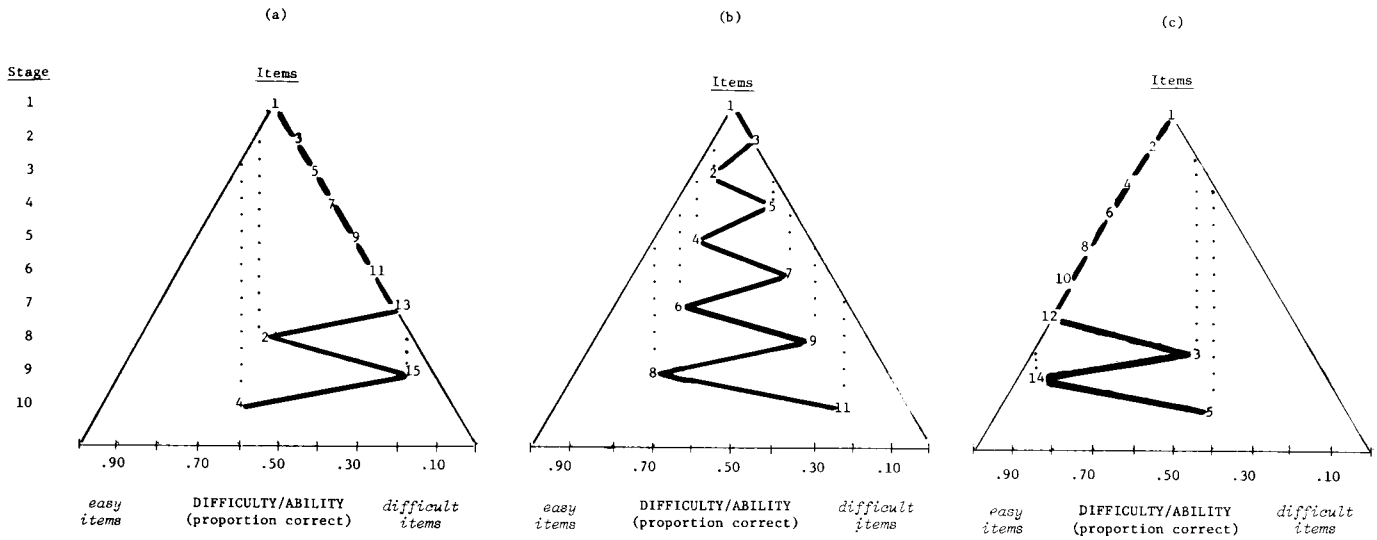


Figure 2

Sample Paths through a Ten-stage Flexilevel Test

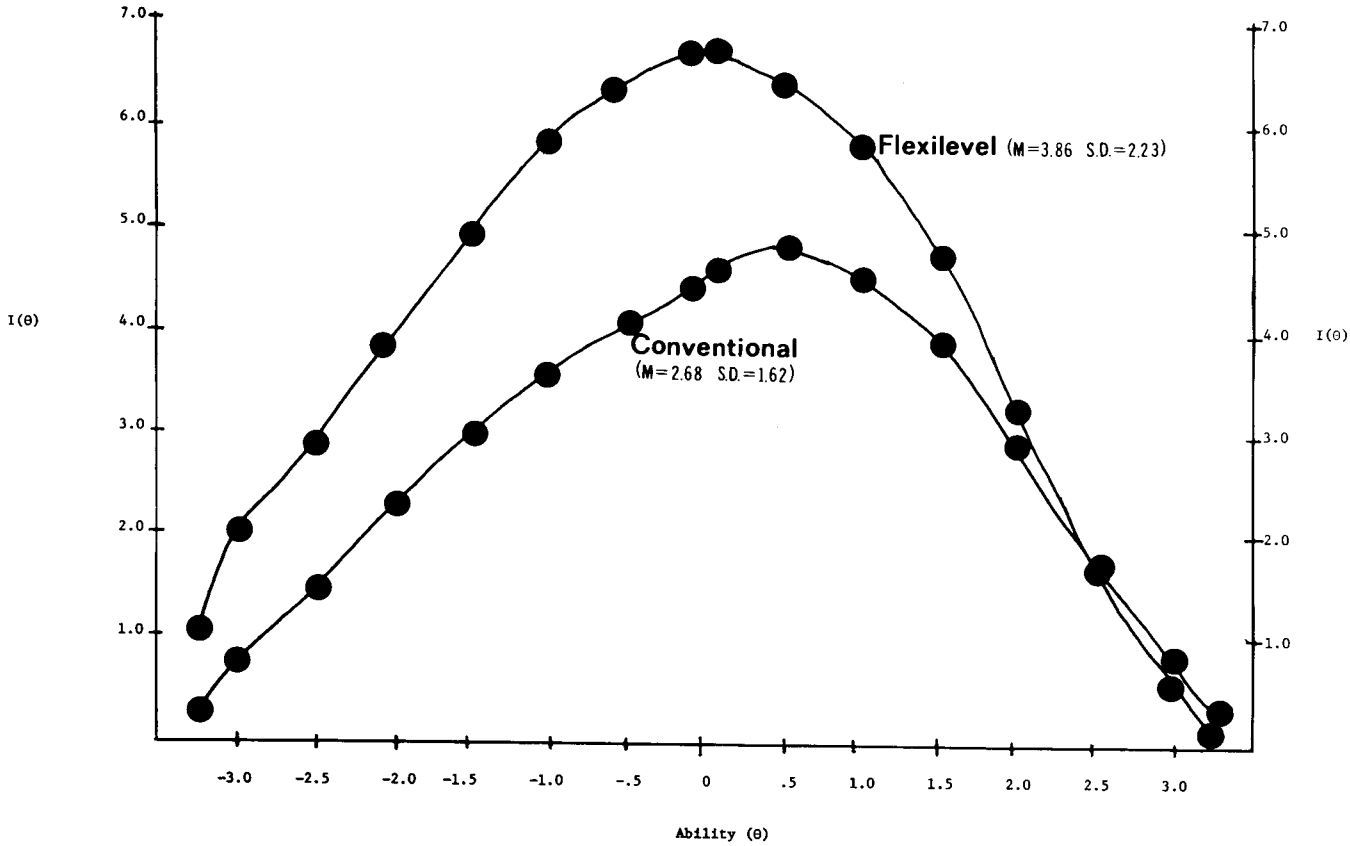


Figure 3

Information functions from flexilevel and conventional tests (N=100 at each of 16 levels of ability)

A comparison of the results from the computer simulation study and the live-testing study showed differences in the test-retest reliabilities. This result was expected because of the memory effects in live testing. There were also differences between the two studies in the shapes of the generated score distributions. These differences demonstrated that the simulation model was not yet adequate enough to reflect the results of live testing and that it needs some revision so that it will enable us to extrapolate from live testing through computer simulation and back to live testing.

Another interesting result from this simulation study relates to the methodology of computer simulation itself. The design of the study was one in which we repeated the computations for a hundred samples of a hundred subjects each in order to study the sampling distribution of the simulation results. This was done to examine the generality of findings from computer simulation studies which use 100 or fewer simulated subjects (e.g., Jensema, 1974; Urry, 1971). We found that estimates of validity, the correlation of generated ability with estimated ability, based on samples of 100, ranged from .87 to .95, with a mean of .91. In certain inter-strategy comparisons different conclusions about the relative utility of a testing strategy might be drawn based on validities of .87 or .95. Thus, simulation studies should be based on samples of more than 100 in order to arrive at stable conclusions.

Two-stage testing. Figure 4 shows a computer report from what we have called a continuous second-stage two-stage test. This adaptive testing procedure was developed by Brad Sympson of our research staff; we later discovered that Fred Lord had independently developed the same testing procedure. In Fall 1975 we tested a number of college students on this continuous second-stage test.

The major problem with two-stage tests as they have been used in the past (Weiss, 1974) is that of routing errors made in branching from the routing test to the measurement test because of errors of measurement in the routing test. To solve this problem, we developed a measurement test stage which consists of a number of very short measurement tests. The example shown in Figure 4 used a 14-item routing test and 25 4-item measurement tests, each at a different level of difficulty. Using this adaptive testing procedure, when an individual completes

REPORT ON TWO - STAGE TEST

NAME: DATE TESTED: 74/12/ 3 I.D. NUMBER:

14 ITEM ROUTING TEST

X = CORRECT 0 = INCORRECT . = ITEM NOT ADMINISTERED

```
I-----------------------------------------------------------I
I                                                           I
I   X   X   X   X   X   X   X   X   X   X   0   0   0   0   I
I                                                           I
I-----------------------------------------------------------I
(EASY)          > > > > > > INCREASING DIFFICULTY > > > > > >          (HARD)
```

ESTIMATED ABILITY FROM ROUTING TEST: 1.4 MEASUREMENT TEST ENTRY POINT: 18

100 MEASUREMENT TEST ITEMS
(36 ITEMS ARE ADMINISTERED)

DIFFICULTY LEVELS:

```
  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
I-----------------------------------------------------------------------------I
I                                                                             I
I  .  .  .  .  .  .  .  .  .  .  .  .  .  X  0  X  X  0  X  X  X  0  .  .  .  I
I                                                                             I
I  .  .  .  .  .  .  .  .  .  .  .  .  .  X  X  X  0  X  X  X  0  0  .  .  .  I
I                                                                             I
I  .  .  .  .  .  .  .  .  .  .  .  .  .  X  X  X  X  X  0  X  X  0  .  .  .  I
I                                                                             I
I  .  .  .  .  .  .  .  .  .  .  .  .  .  0  X  X  X  X  X  X  0  X  .  .  .  I
I                                                                             I
I-----------------------------------------------------------------------------I
  -3.0   -2.5   -2.0   -1.5   -1.0   -.5    0    .5   1.0   1.5   2.0   2.5   3.0
(EASY)                                                                  (HARD)
```

FINAL ABILITY ESTIMATE: 1.30

PERCENTILE SCORE : 90.26

OTE: NORM GROUP FOR ABILITY ESTIMATE AND PERCENTILE SCORE CONSISTED OF UNIVERSITY OF MINNESOTA UNDERGRADUATE AND GRADUATE STUDENTS

Figure 4

REPORT ON TWO-STAGE TEST

the routing test his score is determined and that score is used to choose an appropriate measurement test. Then, to reduce routing errors, a number of measurement tests on either side of the chosen measurement test are also administered to the individual. In the example shown in Figure 4, the individual's score on the routing test estimated his ability at 1.4 standard deviations above the mean. Consequently, the most appropriate measurement test was estimated to be number 18, which had items at difficulty about 1.4 standard deviations above the mean. But, to compensate for possible errors of measurement in the routing test, he was also administered items in measurement tests 14 through 17 and 19 through 22, for a total of 36 measurement test items. These items varied in difficulty from about .25 to 2.25 S.D.'s on the difficulty continuum.

Following a design that we have used in a number of other studies, we did a test-retest live-testing study with this continuous second-stage two-stage test (in which each testee completed 50 items) and a 50-item conventional peaked test, over about a five-week period, with 104 testees. To keep scoring method the same for both testing strategies, maximum likelihood scoring was used for both the two-stage and conventional test.

The study was designed also to equate the two testing procedures for 1) item discriminations; 2) memory effects; and 3) number of items. Memory effects were equated by first determining the number of items each individual repeated on retest of the two-stage test. Then the retest of the conventional test was structured to have the same number of repeated items by inserting the appropriate number of new items.

The test-retest correlation was .94 for the continuous two-stage test and .66 for the equivalent conventional test. Since the difference in stabilities was considerably larger than found in our previous studies of conventional vs. adaptive testing strategies (e.g., Betz & Weiss, 1973, 1975; Larkin & Weiss, 1974), we carefully examined the distribution of conventional test scores derived from the maximum likelihood scoring. Six testees were found with very low ability scores, apparently due to guessing on the conventional test. Data for these testees were eliminated and the test-retest correlations were recalculated. The stability correlation for the two-stage test was .93 and the conventional test .89. This result was similar to that obtained in other comparisons of conventional and adaptive strategies, showing a higher test-retest correlation for the adaptive test than for the peaked conventional test. This result was obtained when both testing strategies were equated for item discriminations and memory effects.

Stradaptive ability testing. The stradaptive testing strategy (Weiss, 1973) is based on a series of peaked tests, each one differing in terms of difficulty. Figure 5 shows the distribution of item difficulties for a hypothetical stradaptive test. In Figure 5 there are nine strata, each of which is a peaked test peaked at a different level of difficulty.

Figure 6 shows an example of an individual moving through a stradaptive test. Testing begins with an item at some point on the difficulty continuum; the entry point is estimated by prior information about the testee. The individual shown in Figure 6 began with the first item at stratum 5, an item of average difficulty. Since he answered that item correctly, he was administered the first item at stratum 6, which consisted of slightly more difficult items. Following the same branching rule—a more difficult item is administered following a correct response, and a less difficult item following an incorrect response—the stradaptive test continues until the termination criterion is reached. The test is terminated when a stratum is identified at which the individual is responding at or below chance level (i.e., 20% or less correct) based on a minimum of five items administered at that stratum. The individual shown in Figure 6 answered five items at stratum 8 and none of them were answered correctly. Consequently the test was terminated since further testing was likely to provide little additional information on the testee's ability level.

Scoring of the stradaptive test results in both ability level scores and consistency scores. Ability level scores reflect the individual's position on the ability scale; consistency scores reflect the variation in item difficulties encountered as the individual goes through the stradaptive test. Figure 7 shows the stradaptive test response record for an inconsistent individual. This person started the test with a relatively difficult item at stratum 8 but answered some easy items incorrectly (e.g., items 8 and 26) and some difficult items correctly (e.g., items 1 and 17). The result was a response record which varied widely across six strata. A comparison of the consistency scores for Figure 7 with those of Figure 6 shows the former to be uniformly higher. Thus, the testee depicted in Figure 7 was more inconsistent in his interaction with this item pool than was the individual in Figure 6.

Our live-testing test-retest study of the stradaptive test was based on about 200 subjects. Over an average five-week period the test-retest reliability for the best method of scoring the stradaptive test was .90; the test-retest reliability for a conventional test using the number of items administered on the average in the stradaptive test (28 items) was .86. This result showed about the same difference in favor of the adaptive test as we have obtained with other adaptive testing strategies.

I had hypothesized earlier (Weiss, 1973) that consistency scores should reflect something about the dimensionality that results from an individual's interaction with an item pool. To extend this hypothesis, if an individual is responding unidimensionally his scores should be more reliable than an individual whose interaction with an item pool is multi-dimensional. In operationalizing this hypothesis, consistency scores were used as an indicator of dimensionality, and test-retest stability as an estimate of reliability. Specifically, testees were divided into five sub-groups on the basis of their time 1 consistency scores, and test-retest reliabilities were computed separately for each of the five sub-groups. The results are shown in Table 1 for consistency score 11, the standard deviation of items encountered.

As Table 1 shows, the highest test-retest stabilities were observed for the very high consistency group for all ten methods of estimating ability within the stradaptive test. The clearest pattern emerged for ability score 1. On that score, the stability for the highly consistent testees was .94, and that for the very low consistency group was .65, with stabilities for the intermediate groups decreasing with decreasing consistency. The possible utility of consistency scores as a moderator variable is that it might permit us to make more stable predictions for some groups of individuals (consistent testees) than for others (inconsistent testees). Particularly noteworthy is the test-retest reliability of .98 for the very highly consistent testees on ability scores 8 and 9.

If these results can be replicated over longer periods of time, the consistency score might prove to be a very useful and powerful moderator variable derivable from a stradaptive testing response record. It appears to be powerful because it also moderates the test-retest reliability, but not

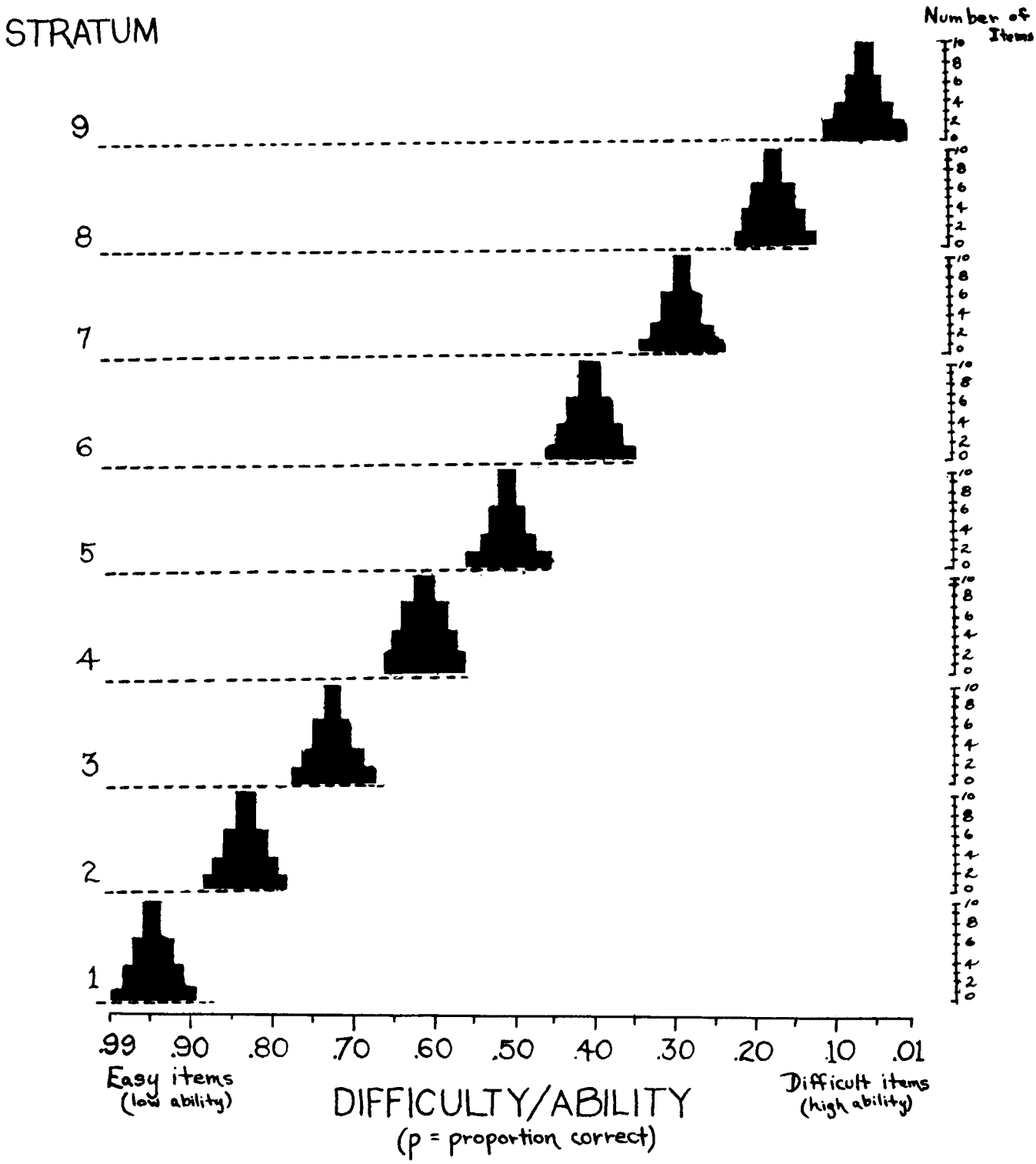
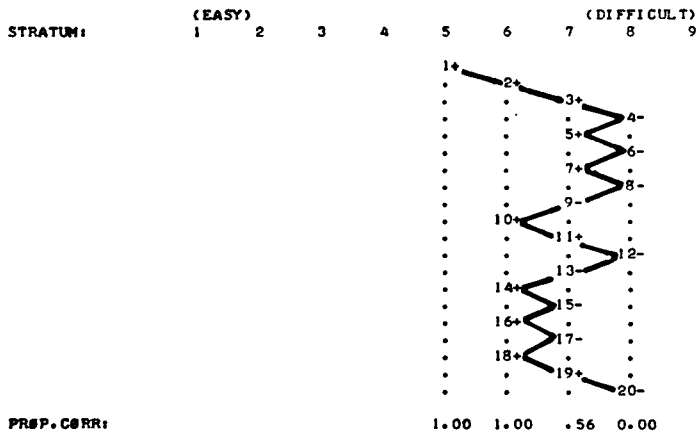


Figure 5

Distribution of Items, by Difficulty Level, in a Stradaptive Test

REPORT ON STRADAPTIVE TEST

ID NUMBER: DATE TESTED: 73/07/12



TOTAL PROPORTION CORRECT= .550

SCORES ON STRADAPTIVE TEST

Ability Level Scores

1. DIFFICULTY OF MOST DIFFICULT ITEM CORRECT= 1.49
2. DIFFICULTY OF THE N+1 TH ITEM= 1.44
3. DIFFICULTY OF HIGHEST NON-CHANCE ITEM CORRECT= 1.49
4. DIFFICULTY OF HIGHEST STRATUM WITH A CORRECT ANSWER= 1.33
5. DIFFICULTY OF THE N+1 TH STRATUM= 1.33
6. DIFFICULTY OF HIGHEST NON-CHANCE STRATUM= 1.33
7. INTERPOLATED STRATUM DIFFICULTY= 1.37
8. MEAN DIFFICULTY OF ALL CORRECT ITEMS= .88
9. MEAN DIFFICULTY OF CORRECT ITEMS BETWEEN CEILING AND BASAL STRATA= 1.28
10. MEAN DIFFICULTY OF ITEMS CORRECT AT HIGHEST NON-CHANCE STRATUM= 1.28

Consistency Scores

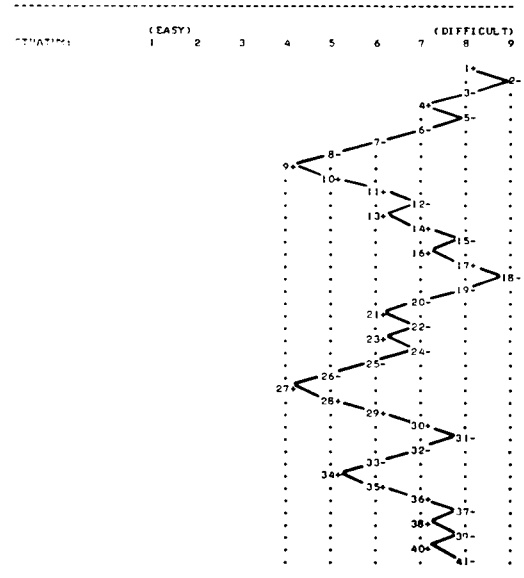
11. SD OF ITEM DIFFICULTIES ENCOUNTERED= .59
12. SD OF DIFFICULTIES OF ITEMS ANSWERED CORRECTLY= .46
13. SD OF DIFFICULTIES OF ITEMS ANSWERED CORRECTLY BETWEEN CEILING AND BASAL STRATA= .18
14. DIFFERENCE IN DIFFICULTIES BETWEEN CEILING AND BASAL STRATA= 1.36
15. NUMBER OF STRATA BETWEEN CEILING AND BASAL STRATA= 1

Figure 6

Report on a Stradaptive Test for a Consistent Testee

REPORT ON STRADAPTIVE TEST

ID NUMBER: DATE TESTED: 73/07/02



TOTAL PROPORTION CORRECT= .488

SCORES ON STRADAPTIVE TEST

Ability Level Scores

1. DIFFICULTY OF MOST DIFFICULT ITEM CORRECT= 1.89
2. DIFFICULTY OF THE N+1 TH ITEM= 1.01
3. DIFFICULTY OF HIGHEST NON-CHANCE ITEM CORRECT= 1.53
4. DIFFICULTY OF HIGHEST STRATUM WITH A CORRECT ANSWER= 2.01
5. DIFFICULTY OF THE N+1 TH STRATUM= 1.33
6. DIFFICULTY OF HIGHEST NON-CHANCE STRATUM= 1.33
7. INTERPOLATED STRATUM DIFFICULTY= 1.36
8. MEAN DIFFICULTY OF ALL CORRECT ITEMS= .72
9. MEAN DIFFICULTY OF CORRECT ITEMS BETWEEN CEILING AND BASAL STRATA= .76
10. MEAN DIFFICULTY OF ITEMS CORRECT AT HIGHEST NON-CHANCE STRATUM= 1.24

Consistency Scores

11. SD OF ITEM DIFFICULTIES ENCOUNTERED= .86
12. SD OF DIFFICULTIES OF ITEMS ANSWERED CORRECTLY= .74
13. SD OF DIFFICULTIES OF ITEMS ANSWERED CORRECTLY BETWEEN CEILING AND BASAL STRATA= .50
14. DIFFERENCE IN DIFFICULTIES BETWEEN CEILING AND BASAL STRATA= 8.64
15. NUMBER OF STRATA BETWEEN CEILING AND BASAL STRATA= 3

Figure 7

Report on a Stradaptive Test for an Inconsistent Testee

TABLE 1

STRADAPTIVE and Conventional test Test-Retest Correlations as a Function of Consistency Score 11 on Initial Testing

		Status on Consistency Score 11				
		Very High	High	Average	Low	Very Low
Mean Consistency Score		.517	.625	.706	.815	1.038
Number of Testees in Interval		27	30	41	43	29
Stradaptive Ability Score:	1	.940	.849	.847	.768	.652
	2	.875	.721	.799	.778	.751
	3	.956	.813	.878	.826	.708
	4	.934	.840	.846	.731	.664
	5	.896	.722	.793	.756	.741
	6	.950	.798	.886	.820	.704
	7	.970	.844	.902	.851	.758
	8	.981	.927	.915	.853	.869
	9	.983	.939	.907	.899	.889
	10	.951	.792	.882	.822	.718
Conventional Test		.979	.890	.918	.826	.878

as systematically, on the conventional test administered at the same time. Table 1 shows a test-retest reliability of .979 on the conventional test for the highly consistent group using the consistency scores derived from the stradaptive test. But consistency scores are not derivable from a conventional test so it is necessary to implement this finding within the framework of the stradaptive testing strategy.

Figure 8 shows a number of "subject characteristic curves," which are derivable from the stradaptive test. These curves, which reflect the individual's consistency of interaction with a stradaptive test, are based on a plot of proportion correct for each individual at each stratum of the stradaptive test. For example, the plot for "William W." shows that he answered all items correctly at both stratum 5 and stratum 6, about half correct at stratum 7 and none correct at stratum 8. Since proportion correct decreases monotonically with increasing item difficulty this individual appears to be interacting with this item pool unidimensionally; William W. is a highly consistent individual. By way of contrast, the subject characteristic curve for "Carol C." does not decrease monotonically, reflecting an inconsistent individual who answers items correctly at a variety of difficulty levels.

To be useful, these subject characteristic curves must be stable across time. To investigate their stability across an average five-week retest interval we computed canonical correlations between proportions correct at initial test and at retest. The complete redundancy analysis showed that 67% of the variance in retest subject characteristic curves was predictable from initial testing. This is equivalent to a squared multiple correlation of .82 for predicting individual proportion correct at Time 2 from a best-weighted linear combination of proportions correct at Time 1. These results imply that subject characteristic curves are reasonably stable and that they may represent a stable trait of the individual. But, certainly, more research is needed.

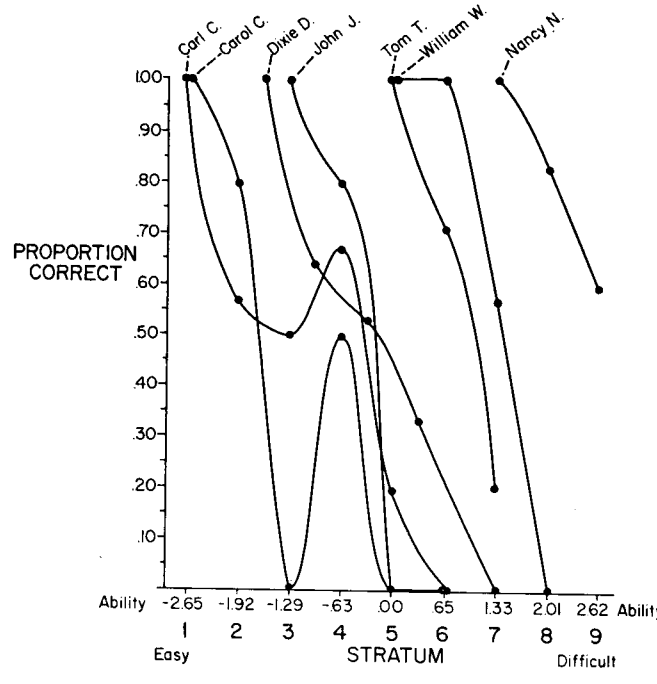


Figure 8

Proportion correct at each stratum, by individual

In addition to this live-testing study of the stradaptive test, we also have some recent data from a computer simulation study. Items with constant discriminations, and difficulties rectangularly distributed between normal ogive difficulty values of −3.33 and 3.33 and grouped into nine equally wide strata were used for the stradaptive test. Items with constant discriminations and with difficulties rectangularly distributed between −.33 and .33 (equivalent to the middle stratum of the stradaptive test) were used for the conventional test. 1000 Ss were generated with abilities in the given interval at each of 13 intervals of θ. Major findings are shown in Figure 9 and Table 2.

Figure 9 shows the information functions for the stradaptive and conventional tests at two different levels of item discrimination. At both levels of item discrimination, the information function for the stradaptive test was more horizontal than that of the conventional test, with the difference more pronounced at the higher level of item discrimination. In confirmation of Lord's theoretical predictions, the conventional test has a higher information function than the stradaptive test at the center of the ability distribution, but the range of superiority diminishes with increasing item discriminations. However, the information function for the stradaptive test increases with ability level, and for the lower discriminating items, the stradaptive test at $\theta \geqslant 2.5$ yields a higher information function than the highest value reached by the conventional test.

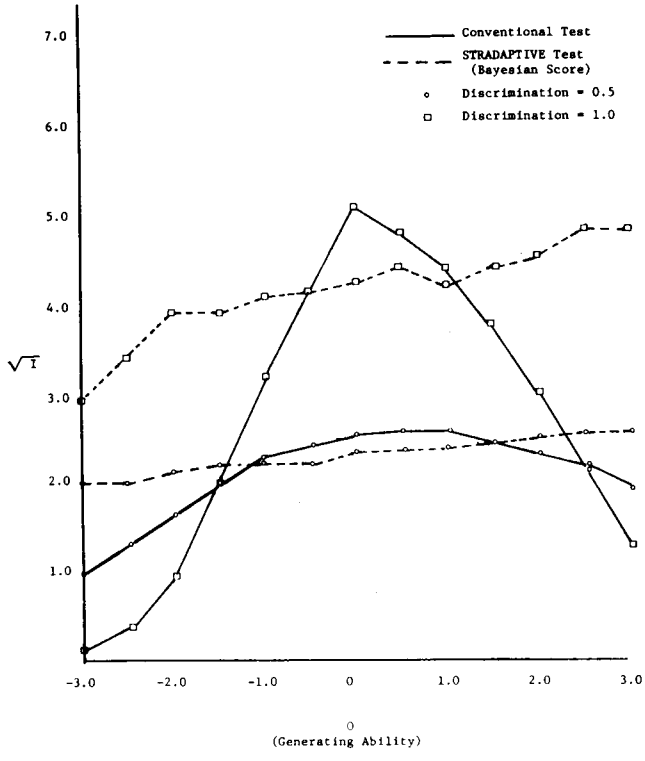


Figure 9

Information Functions for 60-Item Tests

Table 2 shows validities—correlations of ability estimate and generated ability—from the simulation data on conventional and stradaptive tests. Validity correlations are shown as a function of both item discriminations and number of items. These results show a slight superiority in validities for the conventinal tests when item discriminations are low (a=.5), and there are 40 or fewer items in both tests; a similar result is found for 10-item tests composed of items at a=1.0. In all other conditions, the stradaptive test yields higher validity, with sizable differences appearing as number of items increases and discriminations increase. For 60-item tests at a=2.0, the validity of the stradaptive test was r=.989, while the conventional test validity was only .926.

Thus, the data from both the live-testing study and the simulation study of stradaptive tests show that the stradaptive test yields scores which are more equi-precise across the ability range, and have higher validities and reliabilities than conventional tests under certain conditions. Further, the stradaptive test consistency scores appear to be powerful moderator variables which may have important practical applications in testing individuals.

Psychological effects of computerized administration. One of the psychological variables that has been unsystematically manipulated in computerized testing studies has been feedback or knowledge of results. In computerized testing we now have the capability to tell an individual whether his answer was correct or incorrect after each item in a test. But it is possible that such immediate knowledge of results might have an effect on test scores. Thus, we designed a pilot study to systematically manipulate feedback and study its effects on test scores.

TABLE 2

Score–Ability Correlations of the Stradaptive Bayesian Score and the Conventional Test Score for Tests of 10 to 60 Items, as a Function of Item Discrimination

No. Items	Discrimination (a) 0.5	1.0	2.0
10			
Strat	.689	.840	.919
Conv	.703	.851	.888
20			
Strat	.798	.918	.963
Conv	.811	.908	.906
40			
Strat	.869	.955	.983
Conv	.887	.938	.918
60			
Strat	.920	.971	.989
Conv	.917	.950	.926

We administered two tests on the computer to a group of inner-city high school students. The group was racially mixed, consisting of both white students and black students. Both a conventional test and a pyramidal adaptive test were administered to each student, and half the group received the conventional test first and half received the adaptive test first. In addition, half the group received feedback after each item and the other half received no feedback after each test item. We analyzed the data for the conventional test only—thus, the dependent variable in this analysis was number correct on the conventional test. The design was a 2x2x2 analysis of variance. The independent variables were 1) race—black and white; 2) feedback—immediate or none; and 3) order—conventional test administered first or second in the pair.

In order to make the feedback relevant to the high school group, we had previously asked a subgroup of students from the same school to generate a set of statements which would, to them, indicate that they answered an item correctly. We used six such statements, in pseudorandom order, including "right on," "that's cool, now try this one." and "all right, how about this one." This was done on the hypothesis that feedback can have an effect only if it is meaningful or relevant to the testee.

The results for the three-way analysis of variance are shown in Table 3. The only significant main effect was for race. Mean scores for the blacks was 17.74 and that for the whites was 27.92, on the 40-item test. Neither order nor

TABLE 3

Mean Test Scores for Blacks and Whites on the 40-item Test in Two Orders and With and Without Feedback

Group	Feedback N	Feedback Mean	No Feedback N	No Feedback Mean	Total Group N	Total Group Mean
Blacks–First	8	26.38	6	13.83	14	21.00
Second	7	13.86	6	14.67	13	14.23
Whites–First	15	26.07	14	30.93	29	28.41
Second	15	30.00	19	25.53	34	27.50
Blacks	15	20.53	12	14.25	27	17.74
Whites	30	28.03	33	27.82	63	27.92
First	23	26.17	20	25.80	43	26.00
Second	22	24.86	25	22.92	47	23.83
Total	45	25.53	45	24.20	90	24.87

3–Way Anova

Source of Variation	DF	Mean Square	F	Est. P
Order	1	105.76	1.36	.25
Race	1	2,013.26	25.84	<.00
Feedback	1	81.74	1.05	.31
Race x Order	1	161.54	2.07	.15
Order x Feedback	1	28.74	.37	.55
Race x Feedback	1	170.40	2.19	.14
Order x Race x Feedback	1	599.46	7.69	<.01
Error	82	77.92		

feedback effects were significant, nor were any of the two-way interactions. The three-way order x race x feedback interaction was significant at $p<.01$.

Figure 10 shows the means for the three-way interaction. As is indicated in Figure 10, under conditions of immediate feedback, when a conventional test was administered first, the mean of the black students (26.38) was not significantly different from the mean of the white students (26.0) who completed the conventional test under the same set of conditions. This result implies, if it can be replicated, that race differences observed in test scores may be a function not of differences in ability but of differences in the psychological effects of the conditions of administration. Although these findings do not completely replicate those of Johnson & Mihal (1973), they do support their general conclusion that conditions of test administration might affect motivational conditions, which in turn reduce race group differences to nonsignificant levels.

There is some data in our results which suggest that the three-way interaction results might be due to motivational effects. In addition to analyzing test scores, we also anlayzed the proportion of items skipped on the conventional test under the two experimental conditions and for the two racial groups. These results showed that blacks skipped more items than whites, in general, but when the conventional test was administered first to the black students and they received feedback, they skipped almost no items. This is also the same set of conditions under which the test scores for the blacks were not significantly different than those of the whites. This appears to be a motivational effect since when the blacks are given feedback the test becomes relevant to them; and when it becomes relevant they can answer the questions just as well as the whites.

Future Plans

Based on these preliminary findings we plan to continue to investigate the nature of feedback effects, and the effects of other psychological variables, on test scores. We also plan to continue to study various branching schemes in an attempt to develop optimal branching schemes which result in maximum reduction in psychometric error at all ability levels. Our general goal, as I indicated earlier, is to explore all aspects of computerized ability testing in an effort to make maximal use of the computer as a vehicle for making each individual's test score as error-free as possible.

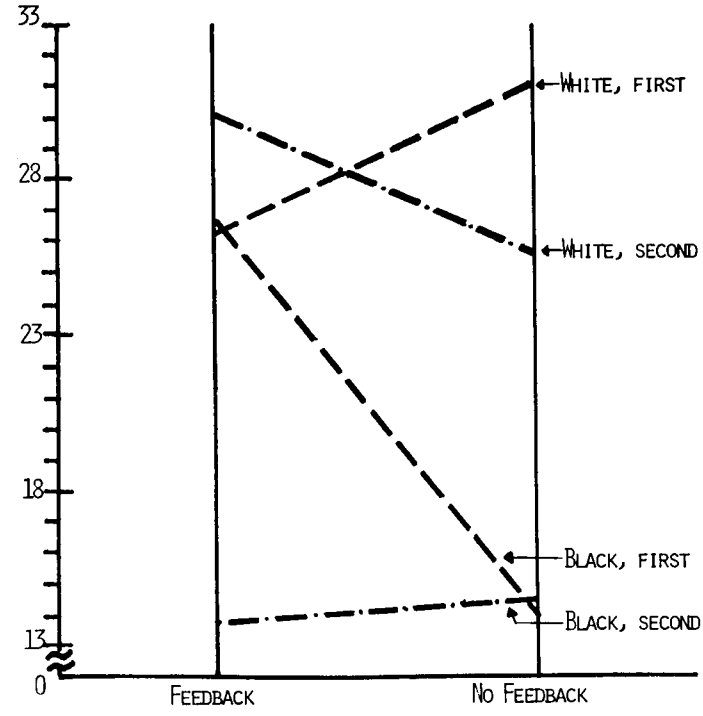


Figure 10

Mean Scores for Blacks and Whites
Completing the 40-item Test First and
Second in Both Feedback Conditions

REFERENCES

Betz, N. E. & Weiss, D. J. An empirical study of computer-administered two-stage ability testing. Research Report 73-4, Psychometric Methods Program, Department of Psychology, University of Minnesota, October 1973.

Betz, N. E. & Weiss, D. J. Simulation studies of two-stage ability testing. Research Report 74-4, Psychometric Methods Program, Department of Psychology, University of Minnesota, October 1974.

Betz, N. E. & Weiss, D. J. Empirical and simulation studies of flexilevel ability testing. Research Report 75-3, Psychometric Methods Program Department of Psychology, University of Minnesota, July 1975.

Jensema, C. J. The validity of Bayesian tailored testing. *Educational and Psychological Measurement,* 1974, *34,* 757-766.

Johnson, D. I. & Mihal, W. M. Performance of blacks and whites in computerized versus manual testing environments. *American Psychologist,* 1973, *28,* 694-699.

Larkin, K. C. & Weiss, D. J. An empirical investigation of computer-administered pyramidal testing. Research Report 74-3, Psychometric Methods Program, Department of Psychology, University of Minnesota, July 1974.

Larkin,K. C. & Weiss, D. J. An empirical comparison of two-stage and pyramidal adaptive ability testing. Research Report 75-1, Psychometric Methods Program, Department of Psychology, University of Minnesota, February 1975.

Lord, F. M. The self-scoring flexilevel test. *Journal of Educational Measurement,* 1971, *8* 147-151. (a)

Lord, F. M. A theoretical study of the measurement effectiveness of flexilevel tests. *Educational and Psychological Measurement,* 1971, *31,* 805-813. (b)

Lord, F. M. Tailored testing, an application of stochastic approximation. *Journal of the American Statistical Association,* 1971, *66,* 707-711. (c)

Owen, R. J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association,* 1975, in press.

Urry, V. W. Individualized testing by Bayesian estimation. Research Bulletin 0171-177. Seattle: Bureau of Testing, University of Washinton, 1971.

Weiss, D. J. The stratified adaptive computerized ability test. Research Report 73-3, Psychometric Methods Program, Department of Psychology, University of Minnesota, September 1973.

Weiss, D. J. Strategies of adaptive ability measurement. Research Report 74-5, Psychometric Methods Program, University of Minnesota, December 1974.

Weiss, D. J. & Betz, N. E. Ability measurement: conventional or adaptive? Research Report 73-1, Psychometric Methods Program, Department of Psychology, University of Minnesota, February 1973.